Monte Carlo approach, is where you run a system a significant number of episodes with quasi-random inputs. The desired outputs are then measured, analysed or learned from directly, by the agent. It improves the policies and values every episode, thus it can only be applied to episodic tasks that eventually terminate, as its learning approach is based off averaging sample returns. In policy iteration, we can use Monte Carlo techniques to both, evaluate a policy and also improve a policy based off empirical simulations.

Monte Carlo approaches however, are computationally slow and expensive. The mean has a convergence of and convergence of covariance is . Therefore, for error to be zero, would need to be infinite but since we always have finite , Monte Carlo estimates are never exact. However typically we work with generalised policy iterations which can work with inexact values.

It can be better than policy and value iteration algorithms because of its real-time learning characteristics. It uses model-free learning, meaning it does not need to have complete prior knowledge of the environment. Furthermore, it can be more efficient to run because it does not need to travel to every single state and so, when calculating policies, will only need to update the states that it has visited which tends to be the most common/important. Unlike policy and value iteration which are model-based and need to sample over an entire state space to specify a policy.